A correlation analysis of game statistics and season results in Major League Soccer

Ana C. Sinicariello

The Ohio State University

Advisor: Dr. Brian A. Turner

Abstract

The aim of this study was to analyze statistical data collected from all games (n = 340) in the 2016 Major League Soccer regular season and discover which statistics are correlated to total season points. Statistical data for 37 variables was collected from the regular season and organized as averages by team. Statistical data for 35 variables was collected from the regular season for the 10 field players with the highest season minutes and organized by player and as averages by team. These variables were then analyzed to determine which were correlated to total season points. The statistics analyzed included crosses, shots on target, yellow cards, aerial duels won, shots against, key passes, offside, player height, etc. The variable of season points refers to the total points a team earns throughout the season by collecting 3 points for each win, 1 point for each tie, and 0 points for each loss. Having the highest amount of total season points is desirable. Multiple team statistics were found to be correlated to total season points such as yellow cards per game (p = 0.0095) and aerial duels won per game (p = 0.0213). Multiple player statistics were found to be correlated, such as average total season minutes (p = 0.00351) and average total assists (p = 0.0610). Besides this, shots on target was found to be significant, but total shots was not. This study aimed to determine the game statistics both by team and by player that are most strongly correlated to total season points.

Introduction

Soccer is known as the international sport; however, North America seems to fall behind the rest of the world in level of play, support of the sport, and technological advancements (Bischoff, 2015). Major League Soccer (MLS) teams are slowly joining in on the statistical revolution of many soccer teams from other nations. The aim of this study was to analyze statistical data collected from all games (n = 340) in the 2016 MLS regular season and discover which statistics were correlated to total season points. The statistics collected were organized by team averages, by each team's 10 field players with the highest season minutes, and by team averages for the same 10 players. Single linear regression and multiple linear regression techniques were used to determine which statistics were correlated to higher total season points. These correlations were then interpreted for further understanding and discussion. This study looks to increase the amount of research and knowledge of Major League Soccer while recognizing that the game of soccer is unique and unpredictable, and cannot be fully reduced to numbers.

Literature Review

The world's favorite sport: a daunting, yet accurate title for the game of soccer (Bischoff, 2015). Along with this large fan base comes a long and distant history of where it all began and how it all started. October 6, 1863 marks the date when Rugby Football and the Football Association parted paths and became two distinct entities in London, England. It was then that the Football Association came up with the original rules for the sport (Sanders, 2013). After examining the original rules of the game of soccer, it is clear that "they bear no relation to modern football. In fact, they are closer to rugby" (Sanders, 2013, p. 40). Other historians argue "that ancient games such as the Greek *episkyros*, the Roman *harpastum*, the Chinese *tsu chu*" and many others "are ancestors or relative types of modern football" (Giossos, Sotiropoulos, Souglis, & Dafopoulou, 2011, p. 130). However, that date in 1863 represents the creation of the first governing body for soccer, and thus its origin. It is clear that the sport has a long and rich history leading up to its current status of the planet's favorite sport.

Going forward in history, England's Football Association grew, and more and more games were being played each year. As clubs with different rulebooks joined the association, they decided to find compromise and alter the rules. In the late 1860s, the Football Association threw out the original rulebook and adopted a new one based on how most clubs at the time were playing the game (Sanders, 2013). The game as we know it today started to take shape. Moving up to competition on a world stage, soccer became an official sport in the Olympic games in 1900. With soccer's growing success there was a need for an international governing body, so in 1904 the Fédération Internationale de Football Association, better known as FIFA, was founded in Paris (FIFA, 1994). The sport continued to grow throughout the early 20th century. When the football community heard that there wasn't going to be a tournament at the 1932 Olympics, they

decided to start their own competition. Uruguay, not only the winners of the Olympic tournaments in 1924 and 1928, but also an economically thriving nation, stepped up to host the first FIFA World Cup in 1930 (FIFA, 1994). And thus, modern football was born.

Along with competition at the international level, club leagues formed. After the founding of the Football Association in England, the Football League was created in 1888 to embrace the idea of professionalism and "increase commercialization of football clubs" (Buraimo, Simmons, & Szymanski, 2006, p. 29). After this distinction, tension did exist, but the Football League continued to grow. In fact, after just 35 years of existence, "it had 88 members divided into four divisions, all employing professional players" (Buraimo, Simmons, & Szymanski, 2006, p. 30). There were discrepancies over allocation of broadcasting sponsorship revenues, and a spike in interest in soccer with England's performance in the 1990 FIFA World Cup, the top clubs from the country decided to create a new league, the Football Association Premier League. The FA Premier League is what is known as the English Premier League today, one of the most lucrative and popular soccer leagues in the world (Buraimo, Simmons, & Szymanski, 2006).

In the world of soccer, North America seems to always be catching up. With the most popular and profitable sports being football, basketball, baseball, and ice hockey, soccer falls short. Although the sport is growing today, soccer in North America does not have as rich of a history or support system as England does. The first prominent club league for the United States and Canada was NASL, or the North American Soccer League. Founded out of the conjunction of two smaller leagues in 1967, it started with 17 teams (Strutner, Parrish, & Nauright, 2014). NASL relied heavily on foreign talent to attract fans to attend matches, however this reliance may have also caused the league's demise. NASL was often referred to as the "'retirement

league,' known for attracting famous foreign players whose best playing days had passed" (Strutner, Parrish, & Nauright, 2014, p. 23). This technique attracted fans to see big European talent play, but the level of the league was so low on the international playing field because of older players who cared less about their teams. With international stars' paychecks running high and interest in the sport running low, NASL was dissolved in 1985 (Scott, 2011). Soccer remained dormant for almost a decade, until the United States hosted the 1994 FIFA World Cup. Building off the enthusiasm and excitement around the sport generated by the World Cup, the United States decided to try again and build a new club league in North America. Major League Soccer, also known as the MLS, had its inaugural season in 1996 and set out to differ from its NASL predecessor. According to Scott (2011),

> the major difference between the MLS and its predecessor is that, rather than simply seeking to crudely impose American values on a sport that has different traditions and then expect both domestic and international success, the MLS is embarking on a much subtler, long-term agenda that requires a certain amount of accommodation with, even learning from, non-American values and sources. (p. 833).

Essentially, Scott is explaining that with NASL, North America tried to make its own version of soccer; its own way of playing the game. However, with the founding of the MLS, the new league attempted to embrace the international sport and play the game as the world does. Although this conscious decision has influenced the MLS to grow and become a strong league in North America, the MLS still lags behind its European counterparts in terms of competition.

Since the founding of FIFA in the early 1900s and the first FIFA World Cup in 1930, the rules of the game itself, soccer, have barely changed (Sanders, 2013). But, the way the game is played and coached is constantly adapting. One of the strategies that is gaining reputability and

importance is the use of statistical analysis on teams. Statistical analysis is more prevalent in other sports such as baseball, basketball, and football, where scores are higher and there is more concrete data to run the numbers on. In soccer, games are low-scoring and there are not many calls that stop the game of play (Thibodeau, 2014). So, statistical analysis in soccer takes a bit of deeper thinking. And, as the need for performance analysis in soccer increases, the need for technology to go along with it is increasing as well. Although hand note-taking and video review can function in adding to a team's statistical awareness, the use of technology to aid this process can completely transform the statistical abilities of a team. As noted by Ballesta Castells, García Romero, Fernández García, and Alvero Cruz (2015) in their review, *Current Methods of Soccer Match Analysis*, "the options commercially available on the market are based on a variety of different methods and…the technology is still at an early stage of development" (p. 786). Essentially, although there are new technologies to aid with statistical analysis of soccer, the technologies have a great amount of room for improvement.

Statistical performance analysis of soccer, whether coded through manual or digital avenues, has been used throughout various research studies to analyze the game. For example, in 2010 Randers et al. comprised a study to discover how player activity pattern related to fatigue development during a soccer match (p. 171). They utilized multiple video and photo tracker technologies as well as GPS to aid in analysis. Discovering how fatigue relates to player activity can assist teams in how to train as well as assisting players on deciding how to distribute energy throughout a game. There have always been postulates of how these factors effect a player, but having statistical evidence behind it only strengthens the claim. Other studies have been done to analyze game play. Lago-Peñas and Gómez-López (2014) developed a study to discover how different scores can be predictors of other game factors. For example, they found that "shots on

goal increased approximately by 14% when teams were 1 goal down or the scores were level" (Lago-Peñas & Gómez-López, 2014, p. 781). Statistical analysis such as these can help coaches and players put past games into perspective in order to improve in the future. For example, making a team aware that they play stronger offensively when they are tied or losing, may help players to get themselves in a strong mindset for games. Other studies have looked at game averages, such as shots on goal, offside and sprints. In a study on runs during a game, Andrzejewski, Chmura, Pluta, and Konarski (2015) discovered that "the mean sprinting distance during a soccer match is about 20 m" for Europa League soccer players (p. 48). The researchers chose to use computer software to objectively analyze the data, rather than manual analysis. Although these generalizations cannot necessarily help specific teams in unique ways, they are still able to provide a level of analysis that was previously unknown. Other studies look at specific positions within a team, and analyze statistically from there. For example, Hongyou, Gómez, Lago-Peñas, Arias-Estero, and Stefani (2015) researched the performance levels of goalkeepers based on the level of team they played on. They discovered that elite level goalkeepers achieved higher performance levels than intermediate and low level goalkeepers. They also noticed differences in performance levels based on if the game resulted in a win, loss or tie. Essentially, elite level goalkeepers were playing their best, no matter the score. Studies such as these help immensely with young talent development. Other studies looked at soccer by position, but focused more on qualitative results rather than statistical quantitative ones. Sporis et al. (2012) examined both attacking and defensive positioning playing styles in a qualitative manner. Whether looking at it quantitatively or qualitatively, numerically or verbally, there are multiple ways to analyze the game of soccer.

In the game of soccer, goals are precious. With so few per game, each goal really does count for a lot more than it may in other sports. For this reason, on top of analyzing speed, positioning, and playing styles, analysis of goal scoring bears heavy weight in the world of soccer statistics. Some statistical research focuses on what leads up to the scoring, or concession, of a goal. Looking both at elite national teams competing in the 2010 FIFA World Cup, as well as a number of youth teams, Shafizadeh, Lago-Penas, Gridley, and Platt (2014) found "that constraining the opponent through marking, pressing, movement in space and defensive skills to lose the ball possession frequently could facilitate the execution of effective attack opportunities" (p. 635). Finding these strong offensive avenues is key to the game of soccer. Other studies have analyzed how goals are scored technically, and the propensity of each type of goal. In a research study by Michailidis, Michailidis, and Primpa (2013), all of the goals scored in the 2012 European Championship were analyzed. The study showed that "most of the goals (40.8%) were scored with shot and then with header (27.6%) and with the inner part of the foot (21.1%)" (p. 367). Discovering how most successful goals are scored statistically can assist coaches in planning out what to practice statistically as well. For example, based on this study, European Championship teams would most likely want to practice direct shooting and headers the most, as well as how to best defend them.

These studies, and most statistical soccer studies, are based on FIFA World Cup soccer and European leagues. Comparatively, few studies have focused on Major League Soccer. As discussed previously, Major League Soccer and the English Premier League play the game differently, at different speeds, and with different strategies. Therefore, the analysis of one league does not always transfer to the other. Some studies have focused on Major League Soccer statistical analysis. Schmicker's (2013) analysis of corner kicks from the 2010 MLS season

focused on spatial distribution of the ball on corner kicks that led to goals. The study found "data from 1859 corner kicks with an overall goal rate of 2.2%. A single box directly in the center of the box, 6-9 yards from goal was the only box with significantly higher rates of goals scored (5.0%) than expected" (Schmicker, 2013, p. 70). Another study looked at the effects of playing tactics on scoring opportunities in the MLS. This study found that "match location (p=0.049), match half (p=0.043) and match status (p=0.032) were associated with creating scoring opportunities" (Gonzalez-Rodenas, Lopez-Bondia, Calabuig, Pérez-Turpin, & Aranda, 2015, p. 851). Another study written by the same authors also looked at the association between playing tactics and counterattacks. This study found that "counterattacks with four or more passes were more effective than shorter ones, regardless of the initial defensive pressure" (Gonzalez-Rodenas, Lopez-Bondia, Calabuig, Pérez-Turpin, & Aranda, 2016, p. 737). Based on the studies mentioned, the goal of my study was to further research of game statistics in the MLS and to create knowledge, without forgetting the notion that the game of soccer cannot be reduced to numbers.

Method

Data

In the Major League Soccer 2016 season there were 20 teams who each played 34 games. WhoScored.com, a website run by a team of skilled soccer analysts with computer science and statistics backgrounds, provides raw statistical data for these regular season matches (WhoScored.com). This site compiles and analyzes statistics from Opta, the official provider of detailed data for the MLS. Three setups of raw data were compiled for this project. Setup 1 included 37 variables of game statistics organized by team. Setup 2 included 35 variables of game statistics organized by player. This setup only included the 10 field players from each team with the highest amount of total season minutes. Setup 3 provided the averages of the 10 field players data from Setup 2 organized by team. I decided to only include the data for the 10 field players with the highest season minutes for each team for multiple reasons. First, I thought a contrast between overall team statistics and this setup could be beneficial. Second, statistics for players who had low amounts of playing time can be skewed. For example, if you only play in one game but you score a goal, you will have a high goals per minute ratio, but it is not necessarily statistically accurate. Third, I chose to exclude goalkeepers because their statistics are completely different than that of field players.

Figure 1

Variables for Setup 1:

Total Season Points	Shots Per Game	Possession Percentage Per Game	Pass Success Percentage Per Game
Aerial Duels Won Per Game	Shots Against Per Game	Tackles Per Game	Interceptions Per Game
Fouls Per Game	Offside Per Game	Shots On Target Per Game	Dribbles Per Game
Fouled Per Game	Total Goals From Open Play Situations	Total Goals From Counter Attack Situations	Total Goals From Set Piece Situations
Total Goals From Penalty Situations	Total Goals From Own Goals Situations	Crosses Per Game	Through Balls Per Game
Long Balls Per Game	Short Passes Per Game	Total Cards From Fouls	Total Cards From Unprofessional Play
Total Cards From Dives	Total Cards from Other Situations	Percentage Of Shots From Inside 6 Yard Box	Percentage Of Shots From Inside 18 Yard Box
Percentage Of Shots From Outside Box	Accurate Crosses Per Game	Inaccurate Crosses Per Game	Accurate Corner Passes Per Game
Inaccurate Corner Passes Per Game	Accurate Free Kicks Per Game	Inaccurate Free Kicks Per Game	Yellow Cards Per Game
Red Cards Per Game			

Figure 2

Variables for Setup 2 (by player) and Setup 3 (as averages for players by team):

Total Season Points	Height In Centimeters	Weight In Kilograms	Total Season Minutes
Total Goals	Total Assists	Total Yellow Cards	Total Red Cards
Shots Per Game	Pass Success Percentage	Aerial Duels Won Per Game	Tackles per game
Interceptions Per Game	Fouls Per Game	Offside Won Per Game	Clearances Per Game
Dribbled Past Per Game	Outfielder Block Per Game	Total Own Goals	Key Passes Per Game
Dribbles Per Game	Fouled Per Game	Offside Per Game	Dispossessed Per Game
Bad Control Per Game	Average Passes Per Game	Crosses Per Game	Long Balls Per Game
Through Balls Per Game	Shots From Outside Box Per Game	Shots From Inside Six Yard Box Per Game	Shots From Inside 18 Yard Box Per Game
Total Key Passes Per Game	Long Key Passes Per Game	Short Key Passes Per Game	

The methods focus on determining which statistics are correlated to the variable of total season points. Total season points was defined as the total points earned by a team in the 2016 regular season of Major League Soccer, excluding playoffs. A team earns 3 points for a win, 1 point for a tie, and 0 points for a loss. Teams want to have as high of points as possible, as only the six teams from each of the two conferences with the highest regular season points qualify for the playoffs. The team with the highest regular season points wins the Supporter's Shield. For the 2016 regular season of Major League Soccer, the maximum possible points a team could have earned was 102, and the minimum 0. The team earning the highest amount of points was FC Dallas with 60 points. The team earning the lowest amount of points was the Chicago Fire with 31 points. The average points earned by a team was 45.65.

Method 1: Simple Linear Regression

The raw data was downloaded into R, a statistical software, and analyzed (R Core Team, 2017). First, a simple linear regression was run for each of the variables on total season points for all three setups. To do this, total season points needs to approximately follow a normal distribution.

Figure 3

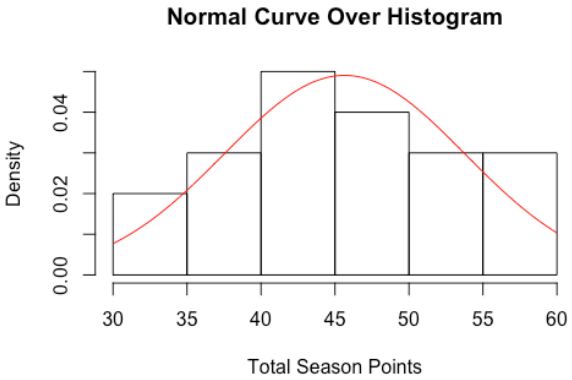
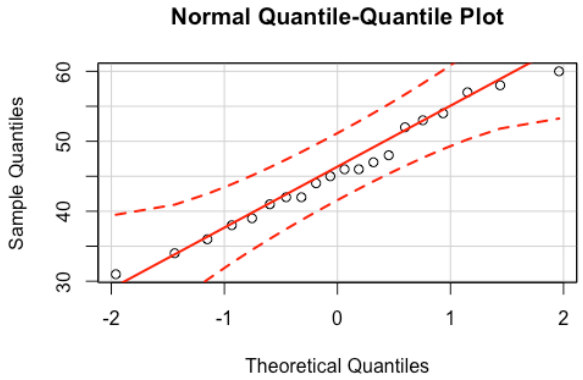


Figure 4



A histogram and a quantile-quantile plot can check if total season points approximately follows a normal distribution. The histogram shows that the data roughly follows a normal distribution curve. The normal quantile-quantile plot plots the sample quantiles against theoretical quantiles for the data. Because this plot resembles a straight line and all of the data points lie within the confidence band, the data is approximately normal. Now that normality has been established, simple linear regression can be run on the data.

Simple linear regression was used to model the correlation of an independent variable and a dependent variable. The null hypothesis states that there is no possible linear function in which the independent variable predicts the dependent variable. The alternative hypothesis states that there is a possible linear function in which the independent variable predicts the dependent variable. In this study, an alpha level of .1 was used, so the values were tested at a 90% confidence level. Assuming the null hypothesis, a simple linear regression was run of each of the

independent variables on total season points, the dependent variable. If the *p*-value for a variable was less than .1, the null hypothesis was rejected and it was accepted that there is a linear correlation between the independent variable and total season points. A positive coefficient indicated that there was a positive linear correlation, or that as the independent variable increased, total season points increased. A negative coefficient indicated that there was a negative linear correlation, or that as the independent variable decreased, the total season points increased.

Method 2: Multiple Linear Regression

Multiple linear regression was used to model the correlation of an independent variable and multiple dependent variables. Adjusted R^2 and Schwarz's Bayesian information criterion (BIC) are two criteria that assist in the choosing of the best fit multiple linear regression models. Essentially, they look at every combination of the given independent variables, and choose a model that best predicts the independent variable. For adjusted R^2, the model with the largest value is the best. For BIC, the model with the lowest value is the best. R software derives the adjusted R^2 and BIC values and presents them in a table (R Core Team, 2017). Once the model with the lowest BIC value and the highest adjusted R^2 value was selected, this model was interpreted as a linear function of independent variables. This equation was the best fit linear predictor of the dependent variable total season points.

Results

Setup 1 - Method 1

A simple linear regression was run for each of the 36 game statistic variables organized by team on the variable of total season points. The game statistic variables returning a positive linear correlation were possession percentage per game ($p = 0.0596$), aerial duels won per game ($p = 0.0213$), shots on target per game ($p = 0.0308$), total goals from open play situations ($p = 0.0426$), total cards received due to fouls ($p = 0.0145$), and yellow cards per game ($p = 0.0095$). The game statistic variables returning a negative linear correlation are shots against per game ($p = 0.0460$), crosses per game ($p = 0.0747$), inaccurate crosses per game ($p = 0.0572$), and red cards per game ($p = 0.0236$).

Setup 2 - Method 1

A simple linear regression was run for each of the 34 game statistics organized by player on total team season points. Note that the player data included is that of the 10 players with the most regular season minutes for each team. The game statistic variables returning a positive linear correlation were total season minutes ($p = 0.0007$), total assists ($p = 0.0919$), total yellow cards ($p = 0.0007$), and aerial duels won per game ($p = 0.0561$). There were no negative linear correlations for this setup.

Setup 3 - Method 1

A simple linear regression was run for each of the 34 game statistics organized by team on total season points. Note that the team data included represents an average for the 10 field players with the most regular season minutes for each team. The game statistic variables returning a positive linear correlation were average total season minutes ($p = 0.00351$), average total goals ($p = 0.0357$), average total assists ($p = 0.0610$), average total yellow cards ($p =$

0.0012), and average aerial duels won per game ($p = 0.0679$). There were no negative linear correlations for this setup.

Setup 1 - Method 2

By evaluating the BIC and adjusted R^2, one can determine a best fit model for the data. The results of those calculations for Setup 1 are displayed in Figure 5.

Figure 5

Model #	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1																										
2										*																
3		*		*																						
4		*		*													*									
5				*													*				*					
6															*		*				*					
7												*			*		*				*					
8										*					*		*				*					

Model #	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	BIC	Adjusted R^2
1									*		-1.681544	0.2807766
2										*	-7.03579	0.4983857
3			*								-14.29589	0.6808489
4			*								-16.881425	0.7424691
5			*						*		-21.054568	0.8071932
6					*	*				*	-24.066126	0.8462339
7					*	*				*	-33.2852	0.9095557
8					*	*	*			*	-38.272544	0.9338058

Figure 5 Key:

Dependent variable: Total Season Points	A: Shots Per Game	B: Possession Percentage Per Game	C: Pass Success Percentage Per Game
D: Aerial Duels Won Per Game	E: Shots Against Per Game	F: Tackles Per Game	G: Interceptions Per Game
H: Fouls Per Game	I: Offside Per Game	J: Shots On Target Per Game	K: Dribbles Per Game
L: Fouled Per Game	M: Total Goals From Open Play Situations	N: Total Goals From Counter Attack Situations	O: Total Goals From Set Piece Situations
P: Total Goals From Penalty Situations	Q: Total Goals From Own Goals Situations	R: Crosses Per Game	S: Through Balls Per Game
T: Long Balls Per Game	U: Short Passes Per Game	V: Total Cards From Fouls	W: Total Cards From Unprofessional Play
X: Total Cards From Dives	Y: Total Cards from Other Situations	Z: Percentage Of Shots From Inside 6 Yard Box	AA: Percentage Of Shots From Inside 18 Yard Box
AB: Percentage Of Shots From Outside Box	AC: Accurate Crosses Per Game	AD: Inaccurate Crosses Per Game	AE: Accurate Corner Passes Per Game
AF: Inaccurate Corner Passes Per Game	AG: Accurate Free Kicks Per Game	AH: Inaccurate Free Kicks Per Game	AI: Yellow Cards Per Game
AJ: Red Cards Per Game			

The lowest BIC value (BIC = -38.2725) and highest adjusted R^2 value (adjR^2 = 0.9338) belong to Model 8. Model 8 represents the best linear model of independent variables to predict total season points. The selected model with coefficients is: total season points = 25.938 + 4.562(shots on target per game) + 1.743(total goals from set piece situations) – 1.997(total own goals) +0.143(short passes per game) – 11.612(accurate corner passes per game) + 5.603(inaccurate corner passes per game) – 1.612(accurate free kicks per game) – 117.001(red cards per game).

Setup 2 and 3 - Method 3

For setup 2 and 3, there was too much interdependence in the data to run a multiple linear regression. The variables are dependent on each other, so it was impossible to find a best fit linear model that includes more than one of them. For example, the data for total long key passes

per game and total short key passes per game is included in the data for total key passes per game. Multiple linear regression works best with independent variables.

Discussion

One point that is important to start with and note is that the game of soccer cannot be reduced to numbers. The use of statistical inference and performance analysis is rapidly increasing in the world of sports and athletics, but doesn't come near to replacing the unpredictability and excitement of games. For soccer, statistical prediction becomes even more difficult as the game is low scoring and has few straightforward data that can be quantified exactly. The beauty of sport is that it cannot be predicted. With that being said, there is still a place for statistics in sports like soccer.

When looking at average game statistics organized by team for the Major League Soccer 2016 season, possession percentage per game, aerial duels won per game, shots on target per game, total goals from open play situations, total cards received due to fouls, and yellow cards per game were positively correlated to total season points and shots against per game, crosses per game, inaccurate crosses per game, and red cards per game were negatively correlated to total season points. These variables do not have a causation relationship, but rather a correlational relationship. That means one cannot conclude that a variable causes the other, but rather that they both follow similar trend lines.

Starting with possession, teams with higher average possession percentage per game, also had higher total season points. This statistic seems intuitive, that a team with the ball more would have more opportunities to score and thus a better chance of winning. Aerial duels are 50/50 chance balls that are in the air. These types of duels are usually two players from opposing teams going for a header. Teams winning more aerial duels also ended with greater total season points. Although total shots on target was correlated, total shots in general was not. That means that teams with higher shots on target had higher total season points, however teams with higher

shots in general did not necessarily have higher total season points. Another statistic that stands out is that total yellow cards is positively, not negatively, correlated to total season points. Teams receiving a greater amount of yellow cards also had greater total season points in general. This could attest to the intensity of a team or how teams who push the boundaries of the rules win more games. Lago-Peñas and Lago-Ballesteros (2011) also found that yellow cards were statistically significant for the Spanish Soccer League 2008-2009 season (Lago-Peñas & Lago-Ballesteros, 2011). Although yellow cards were positively correlated, red cards were negatively correlated to total season points. When a player receives a red card, the team must play with one less player on the field for the rest of the game. Clearly, this is a strong disadvantage for the team, so it makes sense that less red cards would correlate to a stronger overall season performance. On average for the 2016 MLS season, teams received .11 red cards per game, or approximately one every ten games. Red cards are not frequent; however they are still correlated to total season points.

When looking at the statistical data for the 10 players from each team with the highest total season minutes, season minutes, yellow cards, assists, and aerial duels won per game were found significant. Many of these overlap and confirm some correlations found with setup 1. Assists, however, is a new correlation. Players with a higher amount of assists were related to teams with higher season points. This can be intuitive, because more goals means more wins and with more goals there are more assists. However, goals can be scored unassisted. Players with stronger assist records also had greater total season points.

For setup 3, the same statistics as averages were found significant as that of setup 2 with an addition of average total goals for the players. Average total goals was positively correlated to total season points, so the teams who had their most frequent 10 field players scoring the most

goals, also had higher total season points. Going along with this, teams with higher average season minutes for these 10 players also had higher total season points. Teams who put out the same 10 field players more frequently also had better seasons. Having the same 10 field players together was correlated with a stronger season.

The equation found for the multiple linear regression on team statistics can be used to predict season outcomes. The statistics that together best represented total season points are shots on target per game, total goals from set piece situations, total own goals, short passes per game, accurate corner passes per game, inaccurate corner passes per game, accurate free kicks per game, and red cards per game. If a coach can maximize this equation, it is likely that the team will do well in their season.

Implications

Work involving statistical analysis in Major League Soccer will be relevant within the world of soccer. First, my research has the potential to broaden the scope of research completed on the MLS. As the MLS is often a less respected league to its European counterparts, more research is typically done with those leagues. Because all leagues and even teams play the game so differently, research will be distinct and relevant to have from the MLS. Any research completed on the MLS will not only broaden the knowledge of North American club teams, it will also give the MLS a stronger stance in the realm of soccer research.

Coaches can utilize the data to adapt their playing style to see if these statistical correlations would benefit their team. The conclusions found are correlations, not causations, so for example playing the same 10 field players more frequently may not cause higher points, but it might have a positive impact on the team. This can assist in lineup decisions, as well as game time decisions.

Players and the media can also utilize the results of this study. Seeing that yellow cards are correlated to a better season record might inspire players to have a stronger presence in the game and play with higher intensity. Also, they may practice headers and winning aerial duels to a greater extent knowing that it is strongly correlated to season results. The media can reference the correlations found when making their predictions for games and creating numerically backed commentary.

Overall, obtaining qualitative and quantitative performance analysis of game statistics in Major League Soccer regular season matches has the power to assist coaches and players before, during and after matches. With a lack of research in this area for the MLS, league specific data

can help coaches, players, and the media to notice correlations in gameplay. If teams are able to use these correlations to win more games, they will end up with more season points.

Issues and Areas for Further Research

This study opens up the opportunity for further research. First, other statistics can be collected and analyzed. I only looked at approximately 70 unique statistics, but there is the ability to collect more detailed and a greater variety of statistics. Also, I ran into issues with dependent data. If only independent statistics are selected, more regressions and statistical tests can be run on the data. Overall, Major League Soccer is an understudied league. There is room to increase research on set plays, goalkeeper tendencies, shots on goal, substitution techniques, off ball movement and many other areas. Goals are precious in the game of soccer, and many are scored off of set plays that begin when the game is stopped. Because these plays can be practiced and prepared for easily by teams, and because they frequently cause goals, more statistical analysis of set plays can be beneficial for teams. I analyzed mostly attacking statistics and some defensive statistics for this study, but did not analyze goalkeeper statistics. It would be interesting to determine which goalkeeper statistical variables are correlated to higher total season points. Off ball movement is more difficult to study as it is not easily quantified. But, as video and statistical software improve, it will become easier to study the correlation of off ball movement to total season points. As the MLS enters its twenty-second season and continues to increase its level of play and world importance, research for this league will increase as well.

References

Andrzejewski, M., Chmura, J., Pluta, B., & Konarski, J. M. (2015). Sprinting activities and distance covered by top level Europa League soccer players. *International Journal Of Sports Science & Coaching*, *10*(1), 39-50. Retrieved from http://spo.sagepub.com

Ballesta Castells, C., García Romero, J., Fernández García, J. C., & Alvero Cruz, J. R. (2015). Current methods of soccer match analysis. / Métodos actuales de análisis del partido de fútbol. *Revista Internacional De Medicina Y Ciencias De La Actividad Física Y Del Deporte*, *15*(60), 785-802. Retrieved from http://cdeporte.rediris.es/revista/revista.html

Bischoff, S. (2015). Soccer Culture in America: Essays on the World's Sport in Red, White and Blue. *Journal Of Popular Culture*, *48*(3), 608-611. doi:10.1111/jpcu.12288

Bradley, P. S., Lago-Peñas, C., Rey, E., & Sampaio, J. (2014). The influence of situational variables on ball possession in the English Premier League. *Journal Of Sports Sciences*, *32*(20), 1867-1873.

Buraimo, B., Simmons, R., & Szymanski, S. (2006). English football. *Journal of Sports Economics*, 7(1), 29-46. doi: 10.1177/1527002505282911

FIFA. (1994). History of FIFA. Retrieved from http://www.fifa.com/about-fifa/who-we-are/history/index.html

Giossos, Y., Sotiropoulos, A., Souglis, A., & Dafopoulou, G. (2011). Reconsidering on the early types of football. *Baltic Journal Of Health & Physical Activity*, *3*(2), 129-134. doi: 10.2478/v10131-011-0013-5

Gonzalez-Rodenas, J., Lopez-Bondia, I., Calabuig, F., Pérez-Turpin, J. A., & Aranda, R. (2016). Association between playing tactics and creating scoring opportunities in counter-attacks

from United States Major League Soccer games. *International Journal of Performance Analysis in Sport*, *16*(2), 737-752.

Gonzalez-Rodenas, J., Lopez-Bondia, I., Calabuig, F., Pérez-Turpin, J. A., & Aranda, R. (2015). The effects of playing tactics on creating scoring opportunities in random matches from US Major League Soccer. *International Journal of Performance Analysis in Sport*, *15*(3), 851-872.

Hongyou, L., Gómez, M. A., Lago-Peñas, C., Arias-Estero, J., & Stefani, R. (2015). Match performance profiles of goalkeepers of elite football teams. *International Journal Of Sports Science & Coaching*, *10*(4), 669-682. Retrieved from http://spo.sagepub.com

Lago-Peñas, C., & Lago-Ballesteros, J. (2011). Game location and team quality effects on performance profiles in professional soccer. *Journal Of Sports Science & Medicine*, *10*(3), 465-471.

Lago-Peñas, C., & Gómez-López, M. (2014). How important is it to score a goal? The influence of the scoreline on match performance in elite soccer. *Perceptual And Motor Skills*, *119*(3), 774-784. doi:10.2466/23.27.PMS.119c32z1

Michailidis, Y., Michailidis, C., & Primpa, E. (2013). Analysis of goals scored in European Championship 2012. *Journal Of Human Sport & Exercise*, *8*(2), S367-S375. doi: 10.4100/jhse.2012.82.05

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Randers, M. B., Mujika, I., Hewitt, A., Santisteban, J., Bischoff, R., Solano, R., & ... Mohr, M. (2010). Application of four different football match analysis systems: A comparative study. *Journal Of Sports Sciences*, *28*(2), 171-182. doi: 10.1080/02640410903428525

Sanders, R. (2013). How football was born. *History Today*, *63*(10), 40-42. Retrieved from http://www.historytoday.com

Schmicker, R. H. (2013). An application of SaTScan to evaluate the spatial distribution of corner kick goals in Major League Soccer. *International Journal Of Computer Science In Sport (International Association Of Computer Science In Sport)*, *12*(2), 70-79. Retrieved from www.iacss.org

Scott, I. (2011). From NASL to MLS: Transnational culture, exceptionalism and Britain's part in American soccer's coming of age. *Journal Of Popular Culture*, *44*(4), 831-853. doi:10.1111/j.1540-5931.2011.00865.x

Shafizadeh, M., Lago-Penas, C., Gridley, A., & Platt, G. K. (2014). Temporal analysis of losing possession of the ball leading to conceding a goal: A study of the incidence of perturbation in soccer. *International Journal Of Sports Science & Coaching*, *9*(4), 627-636. Retrieved from http://spo.sagepub.com

Sporis, G., Samija, K., Vlahović, T., Milanović, Z., Barisić, V., Bonacin, D., & Talović, M. (2012). The latent structure of soccer in the phases of attack and defense. *Collegium Antropologicum*, *36*(2), 593-603. Retrieved from http://www.collantropol.hr/antropo

Strutner, M., Parrish, C., & Nauright, J. (2014). Making soccer "major league" in the USA and beyond: Major League Soccer's first decade. *Sport History Review*, *45*(1), 23-36. doi:10.1123/shr.2012-0017

Thibodeau, A. (2014). Statistical analysis development in soccer. *Soccer Journal*, *59*(2), 68-72. Retrieved from https://www.nscaa.com/education/resources/soccer-journal

WhoScored.com. (2016). Major League Soccer. Retrieved from https://www.whoscored.com/Regions/233/Tournaments/85/USA-Major-League-Soccer